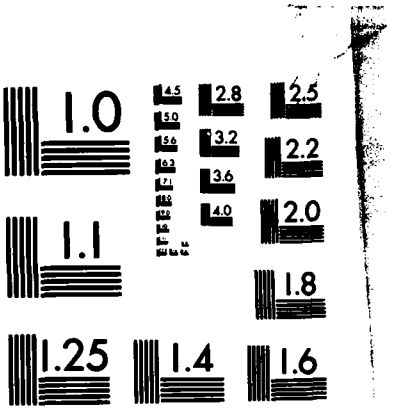
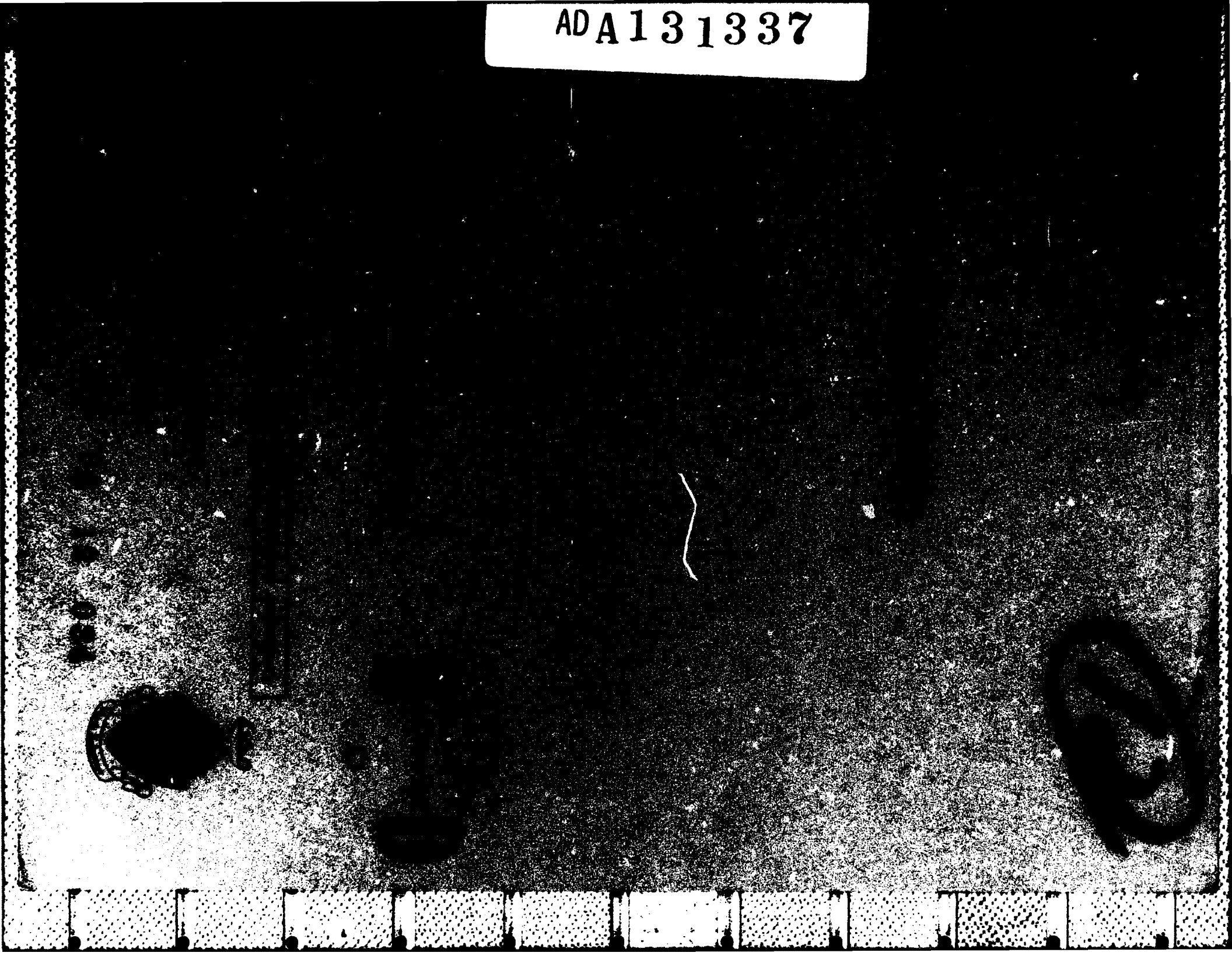
AD-A131 337 MATCHED SURVIVAL ANALYSIS (MSURV)(U) SCHOOL OF AEROSPACE MEDICINE BROOKS AFB TX J E MICHALEK ET AL. MAY 83 USAFSAM-TR-83-15 1/1

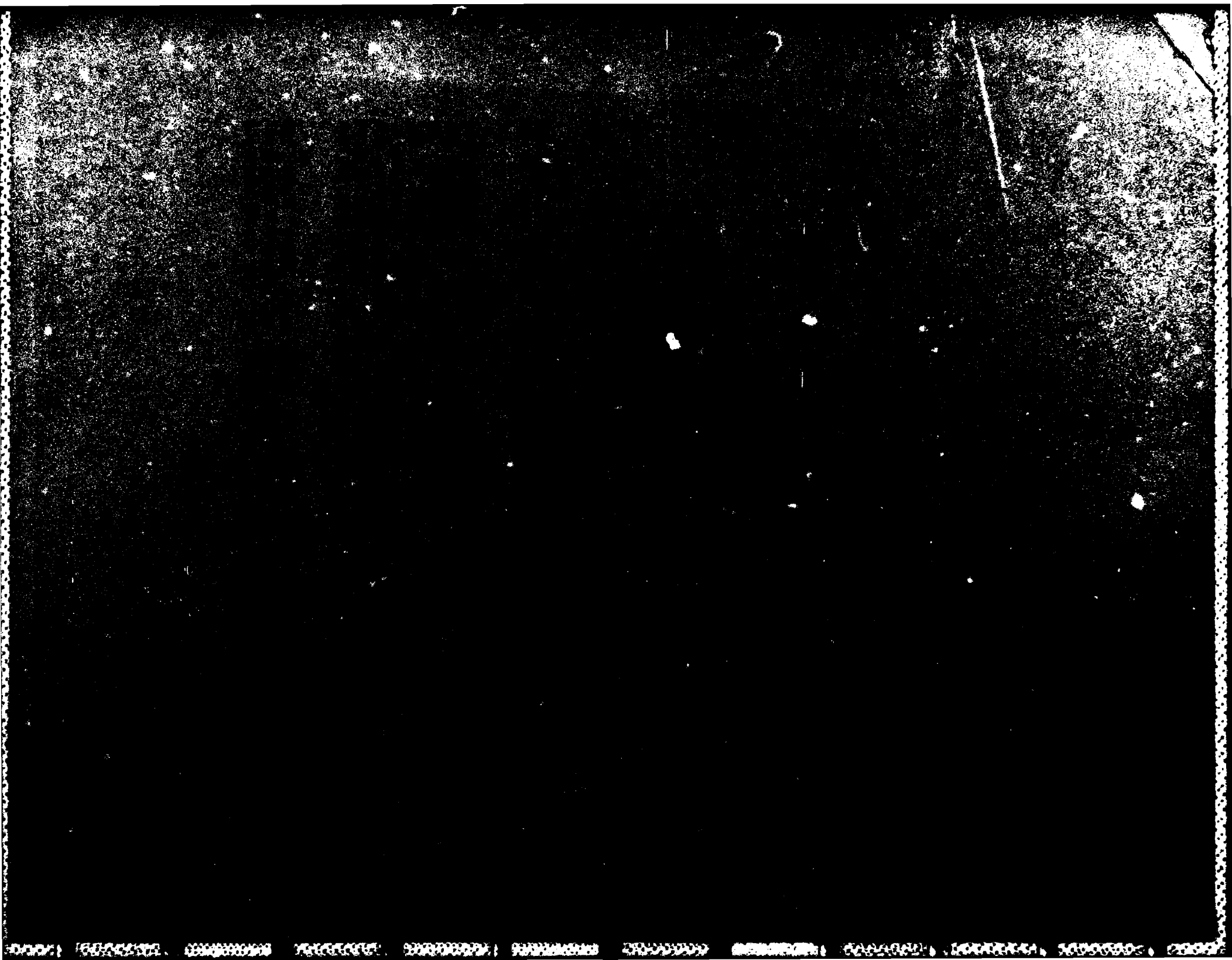
UNCLASSIFIED F/G 12/1 NL

END

FILMED







UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER USAFSAM-TR-83-15	2. GOVT ACCESSION NO. A131337	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) MATCHED SURVIVAL ANALYSIS (MSURV)		5. TYPE OF REPORT & PERIOD COVERED Final Report June 1981 - March 1982 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Joel E. Michalek, Ph.D. Daniel Mihalko, Ph.D. Thomas J. White, M.S.		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS USAF School of Aerospace Medicine (BRM) Aerospace Medical Division (AFSC) Brooks Air Force Base, Texas 78235		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62202F 276700F1
11. CONTROLLING OFFICE NAME AND ADDRESS USAF School of Aerospace Medicine (BRM) Aerospace Medical Division (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE May 1983 13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Daniel Mihalko is Assistant Professor, Department of Mathematics and Statistics, University of Nebraska, Lincoln, Neb. 68588. This report was written while he was a USAF University Resident Research Associate at USAFSAM. Thomas J. White is a Programmer (under contract with OAO Corp., 1222 N. Main Ave., San Antonio, Tex. 78212), in the USAFSAM Biomathematics Modeling Branch.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

arbitrary right censorship
linear rank procedures
matched designs
survival analysis

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report documents a statistical package (MSURV) designed to analyze matched observations which are subject to arbitrary right censorship. The input data are assumed to arise from a matched survival study, in which cases are matched to controls with respect to variables suspected of being confounders. MSURV features: life table output for cases and controls; Kaplan-Meier survival curve estimates, and confidence bands for cases and controls with plotting option; linear rank tests for comparing cases and controls; and a test for comparing cases and/or controls with a life table distribution.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

C O N T E N T S

BLANK PAGE

MATCHED SURVIVAL ANALYSIS (MSURV)

1. INTRODUCTION

The purpose of this statistical package is to analyze matched observations which are subject to arbitrary right censorship. The input data are assumed to arise from a matched survival study in which cases are matched to controls with respect to confounding variables suspected. The input data are therefore assumed to consist of match sets of the form $M = [(t_1, \Delta_1, z_1), (t_2, \Delta_2, z_2), \ldots, (t_m, \Delta_m, z_m)]$

where, for $\nu=1,2, \ldots,m$, t_ν are survival, or response, times,

$\Delta_\nu = 1$, if t_ν is an uncensored observation

$= 0$, if t_ν is censored

and

$z_\nu = 1$, if the subject responding at t_ν is a case

$= 0$, if the subject responding at t_ν is a control.

Censorship is assumed to be due to loss-to-followup, death from causes other than those of interest, or survival to the time of analysis.

The methods in this program are designed to accommodate match sets, of varying length, which have been allocated to several strata. The numbers of case and of control subjects are also allowed to vary from match set to match set. As a special case, this program analyzes match sets arising from a 1-to-R matched design, in which each case is matched to R controls, $R \geq 1$.

In this documentation we describe the MSURV features: life table output for cases and controls; Kaplan-Meier survival curve estimates, and confidence

bands for cases and controls with plotting option; linear rank tests for comparing cases and controls; and a test for comparing cases and controls with a Life Table distribution. The Kaplan-Meier survival curve estimate is programmed from Kalbfleisch and Prentice (Ref. 3: p. 16, eq. 1.10). The confidence band routine is programmed from Hall and Wellner (Ref. 2). The linear rank tests, Michalek and Mihalko (Ref. 4), are match set versions of the log-rank and Wilcoxon tests for censored data. The test for comparing a study population with a Life Table distribution is programmed from Gail and Ware (Ref. 1).

2. LIFE TABLE OUTPUT AND SURVIVAL CURVE ESTIMATES

Letting $t_{(0)}<t_{(1)}<\ldots<t_{(k)}$ denote the distinct ordered uncensored observations, with $t_{(0)}=0$, the following defined numbers are printed at each $t_{(i)}$, i=0,1,2, ...,k:

AT RISK = the number of subjects alive and still in the study at time $t_{(i)}-0$

FAILURES = the number of subjects dying at $t_{(i)}$

CENSORED = the number of subjects censored in $[t_{(i)},t_{(i+1)})$

F(T) = the value of the Kaplan-Meier estimate at $T=t_{(i)}$,

where

$$F(t) = \prod_{i \mid t_{(i)}<t} [(n_i-d_i)/n_i],$$

with n_i = AT RISK, and d_i = FAILURES at $t_{(i)}$.

Two such tables are printed--one for the cases, and one for the controls.

In addition, letting $t^+_{(0)}<t^+_{(1)}<\ldots<t^+_{(L)}$ denote the distinct ordered censoring times, with $t^+_{(0)}=0$, the following defined numbers are, as an option, printed at each $t^+_{(j)}$, j=0,1, ...,L:

AT RISK = the number of subjects in the study at risk of being censored at $t^+_{(j)}-0$

CENSORED = the number of subjects censored at $t^+_{(j)}$

FAILURES = the number of subjects dying in $[t^+_{(j)}, t^+_{(j+1)})$

H(T) = the value of the Kaplan-Meier estimate of the censoring distribution at $T=t^+_{(j)}$.

Two such tables are printed as an option--one for the cases, and one for the controls.

3. LINEAR RANK PROCEDURES

All linear rank procedures in MSURV are of the form:

$$Q = \frac{\sum_{\ell=1}^{n} T_\ell}{(\sum_{\ell=1}^{n} \hat{\sigma}^2_\ell)^{1/2}}, \tag{1}$$

in which T_ℓ is the value of a linear rank statistic T computed on a match set M_ℓ of the form M; $\hat{\sigma}^2$ is an estimate of VAR(T); and $\hat{\sigma}^2_\ell$ is the value of $\hat{\sigma}^2$ computed on M_ℓ. The test T is taken from the class of efficient scores procedures (Ref. 3: p. 146, eq. 6.6):

$$T = \sum_{i=1}^{k} (c_i z_{(i)} + \sum_{j=1}^{m_i} c_i z_{ji}), \tag{2}$$

where, with $t_{(1)}<t_{(2)}<\ldots<t_{(k)}$ denoting the ranked times of the k, $k\leq m$, uncensored observations in the match set M; and, with $t_{(0)}=0$ and $t_{(k+1)}=\infty$, m_i is the number of censored observations in $[t_{(i)}, t_{(i+1)})$, $z_{ji}=1(0)$, if the jth censored subject in $[t_{(i)}, t_{(i+1)})$ is a case (control), $i=1,2,\ldots,k$, $j=1,2,\ldots,m_i$; and $z_{(i)}=1(0)$, if the subject with uncensored response time $t_{(i)}$ is a

case (control), i=1,2, ...,k. The scores c_i and C_i, i=1,2, ...,k, are functions of the observed pattern of censoring and the density of the actual response time (Ref. 5: eq. 17).

Formally, the observed response time, t, is the minimum of an actual response time, t°, and a censoring time u, t=min(t°,u). The hypothesis to be tested:

$$H_0: f_\ell(t_1^\circ, t_2^\circ, \ldots, t_m^\circ) \text{ is symmetric in its arguments,}$$

$$\ell=1,2, \ldots,n,$$

where $f_\ell(t_1^\circ, t_2^\circ, \ldots, t_m^\circ)$ is the joint density of the actual response times in the ℓth match set, ℓ=1,2, ...,n. If the observations within match sets are assumed independent, H_0 reduces to

$$H_0': F_{0\ell}(t^\circ)=F_{1\ell}(t^\circ), \ \ell=1,2, \ldots,n,$$

where $F_{1\ell}(t^\circ)$ and $F_{0\ell}(t^\circ)$ are the distribution functions of the case and control populations for M_ℓ, ℓ=1,2, ...,n. The tests of the form Q are designed to test H_0 or H_0'. Programmed in MSURV are four special cases of Q, corresponding to two versions of $\hat{\sigma}^2$ and two sets of scores, c_i and C_i, i=1,2, log-rank and Wilcoxon scores.

The two variance estimates are:

$$\hat{\sigma}_B^2 = \sum_{i=1}^{k} a_i^2 p_i(1-p_i), \tag{3}$$

where $a_i=c_i-C_i$ and $p_i=n_{1i}/n_i$, with n_{1i} being the number of cases at risk and n_i being the total number of subjects at risk at $t_{(i)}-0$, i=1,2, ...,k, and

$$\hat{\sigma}_U^2 = T^2. \tag{4}$$

The logrank version of T has scores:

$$c_i = \sum_{\kappa=1}^{i} n_\kappa^{-1} - 1, \; C_i = \sum_{\kappa=1}^{i} n_\kappa^{-1}, \; i=1,2, \ldots,k, \tag{5}$$

and the Wilcoxon version of T has scores:

$$c_i = 1-2 \prod_{\kappa=1}^{i} \frac{n_\kappa}{n_\kappa+1}, \; C_i = 1- \prod_{\kappa=1}^{i} \frac{n_\kappa}{n_\kappa+1}, \; i=1,2, \ldots,k. \tag{6}$$

The binomial variance estimate for the logrank test is, therefore,

$$\sum_{i=1}^{k} p_i(1-p_i) \tag{7}$$

and the binomial variance estimate for the Wilcoxon test is:

$$\sum_{i=1}^{\kappa} \left(\prod_{\kappa=1}^{i} \frac{n_\kappa}{n_\kappa+1} \right)^2 p_i(1-p_i). \tag{8}$$

MSURV calculates the following two logrank and two Wilcoxon tests on each data set, $M_1, M_2, \ldots, M_n$:

LU = a logrank test of the form Q with each T of the form (eq. 2) with scores (eq. 5) and $\hat{\sigma}^2=\hat{\sigma}^2_U$, given by (eq. 4)

LB = a logrank test of the form Q with each T of the form (eq. 2) with scores (eq. 5) and $\hat{\sigma}^2=\hat{\sigma}^2_B$, given by (eq. 7)

WU = a Wilcoxon test of the form Q with each T of the form (eq. 2) with scores (eq. 6) and $\hat{\sigma}^2=\hat{\sigma}^2_U$, given by (eq. 4)

WB = a Wilcoxon test of the form Q with each T of the form (eq. 2) with scores (eq. 6) and $\hat{\sigma}^2=\hat{\sigma}^2_B$, given by (eq. 8).

Each test is coded to accommodate tied data. When ties among the uncensored observations are present, the scores are first calculated as if no ties existed, and then the average of the scores for the tied observations are used for each tied observation. The probability that a standard normal variable will exceed the observed value of each statistic, the p-value, is printed for each statistic. Data with cases dying before the controls will produce large values of the statistics. Michalek and Mihalko have discussed these and other linear rank tests for matched designs (Ref. 4).

4. THE GAIL AND WARE TEST

The following procedure is derived from M. H. Gail and J. H. Ware (Ref. 1). The Gail and Ware procedure tests the null hypothesis of equality of a study population survival distribution and a known survival distribution given in the form of a life table, under the assumption that the study hazard function is proportional to the tabled hazard function. The alternatives are that: (a) the study survival time is stochastically less than the tabled values· or (b) the study survival time is stochastically greater than the tabled values.

The input consists of two data sets: a sample of survival data, and a published life table. The survival data is of the form:

$$(x_1,\Delta_1),\ (x_2,\Delta_2),\ \ldots,\ (x_N,\Delta_N),$$

where x_ℓ, $x_\ell \geq 0$, is the age at death or censoring of the ℓth subject, $\ell=1,2,\ldots,N$, and:

$$\Delta_\ell = 1, \text{ if } x_\ell \text{ is an age at death}$$
$$= 0, \text{ if } x_\ell \text{ is an age at censoring, } \ell=1,2,\ldots,N,$$

where censoring may be due to: loss-to-followup; death from causes other than those of interest; or survival to the time of the analysis.

The published life table is of the form:

Age interval	Number surviving	$\hat{q}$
$[a_0, a_1)$	m_1	$\hat{q}_0$
$[a_1, a_2]$	m_2	$\hat{q}_1$
	. . .	
$[a_I, a_{I+1})$	m_I	$\hat{q}_I$

where, for $j=0,1, \ldots,I$,

$$q_j = P[\text{dying in } [a_j, a_{j+1}) \mid \text{survival up to age } a_j]$$

and, with $a_0=0$ and $a_{I+1}=\infty$,

$$\hat{q}_j = (m_j - m_{j+1})/m_j, \; j=0,1, \ldots,I.$$

Letting T_L denote survival time in the life table population and T_S denote population and T_S denote survival time in the study population, we want to test, under the proportional hazards assumption:

$$H_0: P(T_L \geq t) = P(T_S \geq t)$$

$$\text{versus } H_1: P(T_L \geq t) > P(T_S \geq t),$$

$$\text{or } H_2: P(T_L \geq t) < P(T_S \geq t).$$

To this end, we define, for j=0,1, ...,I:

n_j = number of study subjects entering $[a_j,a_{j+1})$

w_j = number of study subjects censored in $[a_j,a_{j+1})$

d_j = number of study subjects dying in $[a_j,a_{j+1})$

s_j = number of study subjects surviving $[a_j,a_{j+1})$,

Define the life-table hazard-function estimate as:

$$\hat{h}(t)\delta = (-1/\delta)\log(1-\hat{q}_j),\ a_j \le t < a_{j+1},$$

where we have assumed that $a_{j+1}-a_j=\delta$, j=0,1, ...,I-1. The expected number, $\hat{e}_j$, of study deaths in $[a_j,a_{j+1})$ is estimated by:

$$\hat{e}_j = n_j\hat{h}(a_j)\delta[1-\hat{h}(a_j)\delta/2-w_j/2n_j].$$

The deviation between the observed number of study deaths, d_j, in the interval $[a_j,a_{j+1})$ and the estimated expected number, $\hat{e}_j$, is simply

$$\hat{D}_j = d_j-\hat{e}_j,\ j=0,1,\ ...,I.$$

The variance of $\hat{D}_j$ is estimated by:

$$\hat{\sigma}_j^2 = [d_j(n_j-d_j)+w_j(n_j-w_j)\hat{h}^2(a_j)\delta^2/4-\hat{h}(a_j)d_jw_j\delta]/n_j.$$

The statistic is:

$$GW = \frac{\sum_{j=0}^{I-1}\hat{D}_j}{\left(\sum_{j=0}^{I-1}\hat{\sigma}_j^2\right)^{1/2}}.$$

The null hypothesis, H_0, is rejected in favor of H_1 at the .05 level of significance when GW≤-1.645; H_0 is rejected in favor of H_2 when GW≥1.645.

The respective p-values are printed. MSURV is coded both to compare cases with the life table using GW, and to make a separate comparison of the controls with the life table.

5. VARIABLE DEFINITIONS

IST	=	stratum number
NCA	=	number of cases in match set
NCT	=	number of cases and controls in match set
AGE(I,N)	=	age at event time, subject I, match set N
X(I,N)	=	event time
DELTA(I,N)	=	censoring indicators
NSTRAT	=	number of strata to be used
ISTRAT(I)	-	contains the numbers of the strata to be used
STNAME	-	contains the names of the strata
NCAS(N)	-	number of cases for the Nth match set
NCON(N)	-	number of controls for the Nth match set
ICNT	=	number of match sets in a stratum
KCAS	=	number of cases in a stratum
KCON	=	number of controls in a stratum.

6. PROGRAM FLOW AND INPUT

The main program loop is over strata. Event times and censoring indicators are passed to subroutine KMSURV in the arrays TS(I) and IDEL(I), I=1,2, ...,ICNT--first for the cases, and then for the controls in the stratum. KMSURV produces tables of survival and: if ICEN≠0, censoring distribution estimates for cases and controls; if IHWBD≠0, a table of confidence bounds for the survival curves for the cases and controls. If IPLOT is not

set equal to zero, survival estimates and confidence bands are plotted. The various statistics and their associated p-values are printed.

The following constitutes the control input to MSURV:

TITLE - up to 80 characters describing the run (20 A4)

The following six control variables are read unformatted:

NSTRAT	=	number of strata to be used
ICEN	-	if not equal to zero, estimates the censoring distribution
IHWBD	-	if not equal to zero, confidence bands are printed
IPLOT	-	if not equal to zero, survival curve estimates and confidence bands are plotted
STBANE	-	the names of the strata, 8 characters each, corresponding to the numbers in ISTRAT, read A8
FMT	-	data format.

Data are read from unit 9 in the following order for each record: stratum number; NCA; NCT; age; event time; and censoring indicator for NCA cases, and for (NCT-NCA) controls.

7. COMPUTATION FORMULAS FOR LINEAR RANK PROCEDURES

7.1 Main Program

The linear rank procedures in MSURV are coded (as follows); and the rows of the data array M are ranked in order of ascending values of t to obtain:

$$\begin{matrix} t_{(1)} & \Delta_{(1)} & z_{(1)} \\ t_{(2)} & \Delta_{(2)} & z_{(2)} \\ \vdots & & \\ t_{(m)} & \Delta_{(m)} & z_{(m)} \end{matrix}$$

The following array is then formed:

$$\begin{matrix} n_1 & d_1 & S_{13} & S_{14} \\ n_2 & d_2 & S_{23} & S_{24} \\ \vdots & & & \\ n_k & d_k & S_{k3} & S_{k4} \end{matrix},$$

in which, for i=1,2, ...,k:

k = number of distinct death times

n_i = number of subjects at risk at $t_{(i)}-0$

d_i = number of deaths at $t_{(i)}$

S_{i3} = sum of the z's for the deaths at $t_{(i)}$

S_{i4} = sum of the z's for the censorings in $[t_{(i)},t_{(i+1)})$.

The vector $N=(N_1,N_2, \ldots,N_D)$ is calculated, where $N_1=n_1$, $N_2=n_1+1$, ...; and $N_{d_1}=n_1+1-d_1$, $N_{d_1+1}=n_2$, $N_{d_1+2}=n_2+1$, ..., $N_{d_1+d_2}=n_2+1-d_2$, $N_{d_1+d_2+1}=n_3$, ..., $N_D=n_k-d_k+1$, where $D=d_1+d_2+ \ldots +d_k$. A subroutine is called to calculate the scores c_υ and C_υ, $\upsilon=1,2, \ldots,D$. The numbers $A_1, A_2, \ldots, A_k$ are calculated by:

$$A_i = d_i^{-1} \sum_{\upsilon=D_{i-1}+1}^{D_i} (c_\upsilon - C_\upsilon), \quad i=1,2, \ldots,k,$$

where

$$D_i = \sum_{j=1}^{i} d_j, \quad i=1,2, \ldots,k,$$

with $D_0=0$. The numerator of Q, given by (eq. 1), is then calculated using:

$$T = \sum_{i=1}^{k} A_i(S_{i3}-d_iR_i/n_i),$$

where

$$R_i = \sum_{j=i}^{k} (S_{j3}+S_{j4}),\ i=1,2,\ \ldots,k.$$

The binomial variance estimate is calculated using:

$$\hat{\sigma}_B^2 = \sum_{i=1}^{k} d_iA_i^2(R_i/n_i)(1-R_i/n_i)[(n_i-d_i)/(n_i-1)]I(n_i=1),$$

in which

$$I(n_i=1) = 0 \text{ if } n_i=1$$

$$= 1 \text{ otherwise, } i=1,2,\ \ldots,k.$$

Note that $I(n_i=1)=0$ is possible only when $i=k$.

7.2 Logrank Score Subroutine

The input is a D-dimensional vector of decreasing nonnegative integers, $N=(N_1,N_2,\ \ldots,N_D)$. The scores are calculated by:

$$c_\nu = 1-\sum_{r=1}^{\nu}(1/N_r),\ C_\nu = -\sum_{r=1}^{\nu}(1/N_r),\ \nu=1,2,\ \ldots,D.$$

7.3 Wilcoxon Score Subroutine

The input is a D-dimensional vector of decreasing nonnegative integers, $N=(N_1,N_2,\ \ldots,N_D)$. The scores are calculated by:

$$c_\nu = 1-2\prod_{r=1}^{\nu}\frac{N_r}{N_r+1},\ C_\nu = 1-\prod_{r=1}^{\nu}\frac{N_r}{N_r+1},\ \nu=1,2,\ \ldots,D.$$

8. REFERENCES

1. Gail, M. H., and J. H. Ware. Comparing observed life table data with a known survival curve in the presence of random censorship. Biometrics 35: 385-391 (1979).

2. Hall, W. J., and J. A. Wellner. Confidence bands for a survival curve from censored data. Biometrika 67: 1, 133-143 (1980).

3. Kalbfleisch, J. D., and R. L. Prentice. The statistical analysis of failure time data. New York: John Wiley, 1980.

4. Michalek, J., and D. Mihalko. Linear rank procedures for matched observations. USAF-SAM-TR-83-16 (In press).

5. Prentice, R. L. Linear rank tests with right censored data. Biometrika 65: 167-179 (1978).